## Last time:

- GD convergence depends on fn proveties
- A single iteration is expensive i.e, requires 1 pass over data

## This lecture:

- Stochastic Grad. Descent
- Convergene and Comparison with GD

---

Property we haven't used yet:

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

i.e, the fn is a finite sum

Simple idea: Cheap "local" updates

- $w_{k+1} = w_k - \gamma \cdot \nabla f_{s_k}(x_k)$

- backprop
- perceptron
- LMS

- $s_k \sim$ i.i.d uniform from $\{1, \dots, n\}$

This step is the answer to a "local" under approx., i.e.,:

$$w_{k+1} = \operatorname{argmin}\left\{ f_{s_k}(w_k) - \langle \nabla f_{s_k}(w_k), w_k - w \rangle + \frac{1}{2\gamma}\|w_k - w\|^2 \right\}$$

## Remarks:

- Simple to implement
- Small memory + computational foot print
- offers simple algorithmic paradigm around which we can build systems

Moreover:

- cost of 1 step: $O(d)$
- $E_{s_k} \nabla f_{s_k}(w) = \frac{1}{n} \sum_i \nabla f_i(w), \ \forall w$

"SGD step = GD on average"

Q: • Does it converge?

- How fast? Comparison to GD

Let's examine convergence properties of SGD:

$$w_{k+1} = w_k - \gamma \cdot \nabla f_{s_k}(w_k)$$
$$s_k \sim \text{uniform}\ \{1, 2, \dots, n\}$$

Assume $f$ is $\lambda$-str. cvx.

- $E\|\nabla f_s(w)\|^2 \le M^2 \quad \forall w \Rightarrow$ Lipschitz

Then,

$$\underbrace{\|w_{k+1} - w^*\|^2}_{\Delta_{k+1}} = \underbrace{\|w_k - w^*\|^2}_{\Delta_k} - 2\gamma\langle \nabla f_{s_k}(w_k), w_k - w^*\rangle + \gamma^2\|\nabla f_{s_k}(w_k)\|^2$$

$$\Rightarrow \mathbb{E}_{s_1 \cdots s_n} \Delta_{k+1} \le \mathbb{E}\Delta_k - 2\gamma\mathbb{E}\langle \nabla f_{s_k}(w_k), w_k - w^*\rangle + \gamma^2 M^2$$

Also, observe that:

$$\mathbb{E}_{s_1 s_2 \cdots s_n} X = \mathbb{E}_{s_1 \cdots s_{k-1} s_{k+1} \cdots s_n} \mathbb{E}_{s_k} X$$

Therefore:

$$\mathbb{E}\langle \nabla f_{s_k}(w_k), w_k - w^* \rangle = \mathbb{E}_{\sim s_k} \mathbb{E}_{s_k} \langle \nabla f_{s_k}(w_k), w_k - w^* \rangle$$

$$= \mathbb{E}_{\sim s_k} \langle \mathbb{E}_{s_k} \nabla f_{s_k}(w_k), w_k - w^* \rangle$$

$$= \mathbb{E}_{\sim s_k} \langle \nabla f(w_k), w_k - w^* \rangle$$

$$= \mathbb{E} \langle \nabla f(w_k), w_k - w^* \rangle$$

Hence,

$$\mathbb{E}_{s_1 \dots s_n} \Delta_{k+1} \leq \mathbb{E}\, \Delta_k - \mathbb{E}\, \gamma \langle \nabla f(w_k), w_k - w^* \rangle + \gamma^2 M^2$$

Due to strong convexity

$$f(w^*) \geq f(w) + \langle \nabla f(w), w^* - w \rangle + \frac{\lambda}{2} \|w - w^*\|^2$$

$$\Rightarrow \langle \nabla f(w), w - w^* \rangle \geq \underbrace{f(w) - f(w^*)}_{\geq 0} + \frac{\lambda}{2} \|w - w^*\|^2$$

$$\Rightarrow \langle \nabla f(w), w - w^* \rangle \geq \frac{\lambda}{2} \|w - w^*\|^2 \quad \forall\, w.$$

(Also true in expectation)

Hence,

$$\mathbb{E}\Delta_{k+1} \leq \mathbb{E}\Delta_k - \gamma\lambda\mathbb{E}\Delta_k + \gamma^2 M^2$$

$$= (1-\gamma\lambda)\mathbb{E}\Delta_k + \gamma^2 M^2$$

$$\leq (1-\gamma\lambda)^2 \mathbb{E}\Delta_{k-1} + \gamma^2 M^2 + (1-\gamma\lambda)\gamma^2 M^2$$

$$\vdots$$

$$\leq (1-\gamma\lambda)^{k+1}\mathbb{E}\Delta_0 + \sum_{i=0}^{k} (1-\gamma\lambda)\gamma^2 M^2$$

Due to $\sum_{i=0}^{\infty} (1-a)^i \leq 1/\alpha \quad \forall\, 0<a<1$

we obtain:

$$\Rightarrow \mathbb{E}\|w_T - w^*\|^2 \leq \underbrace{(1-\gamma\lambda)^T \|w_0 - w^*\|^2}_{\text{Similar to GD}} + \underbrace{\gamma \frac{M^2}{\lambda}}_{\text{Due to "variance"}}$$

We would like the above to be $\varepsilon$

$$\underbrace{(1-\gamma\lambda)^T \|w_0 - w^*\|^2}_{=\varepsilon/2} + \underbrace{\gamma \frac{M^2}{\lambda}}_{\varepsilon/2} = \varepsilon$$

From the second term we get

$$\gamma = \frac{\varepsilon \lambda}{2M^2}$$

From the first term:

$$(1-\gamma\lambda)^T R^2 = \varepsilon/2$$

$$\Rightarrow T \log(1-\gamma\lambda) + 2\log R = \log \varepsilon/2$$

$$\Rightarrow T = \frac{\log \varepsilon/2 - 2\log R}{\log(1-\gamma\lambda)}$$

$$\leq \frac{\log \varepsilon/2 - 2\log R}{-\gamma\lambda}$$

$$\leq \frac{2\log(R/\varepsilon)}{\varepsilon \frac{\lambda}{2M^2}}$$

$$= 4\frac{M^2}{\lambda^2}\frac{\log(R/\varepsilon)}{\varepsilon}$$

## Comparison with GD:

**SGD** on $\lambda$-str. cvx + $M^2$ grad bound

$$T_\varepsilon = O\left(\frac{M^2}{\lambda^2} \log(R/\varepsilon)\right)$$

**GD:** on $\lambda$- str. cvx + $\beta$ smooth:

$$T = O\left(\frac{\beta}{\lambda} \log(R/\varepsilon)\right)$$

**Example:**

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+e^{-y_i\langle w, x_i\rangle}\right) + \frac{\lambda}{2}\|w\|^2$$

- asm :

$$\|x_i\| = O(\sqrt{d}), \quad \|w_0 - w^*\| = O(\sqrt{d})$$

$$\forall w, \ \|w\| \leq O(\sqrt{d}), \quad \lambda = O(1)$$

Then $f(w)$ is
- $O(1)$- str. cvx
- $O(\sqrt{d})$ - Lip
- $O(d)$ - Smooth.
- $M^2 = O(d)$

SGD:

$$T_\varepsilon = O\left(\frac{d}{\lambda^2} \log\left(\frac{d}{\varepsilon}\right) \Big/ \varepsilon\right)$$
$$= O\left(\frac{d}{\varepsilon} \log(d/\varepsilon)\right)$$

GD:

$$T_\varepsilon = O(d \log(d/\varepsilon))$$

But cost of 1 iter of GD

$O(nnz(X))$

cost of 1 iter of SGD $O\left(\frac{nnz(X)}{n}\right)$

$$\Rightarrow \quad \frac{\text{time}(GD, \varepsilon)}{\text{time}(SGD, \varepsilon)} = \frac{O(nnz(A)\, d \log(d/\varepsilon))}{O\left(\frac{nnz(A)}{n} \frac{d}{\varepsilon} \log(d/\varepsilon)\right)}$$

$$= O\left(n \frac{\log(1/\varepsilon)}{\varepsilon}\right)$$

$\Rightarrow$ when $\varepsilon \gg 1/n$ SGD is much faster!

Remark 1:

Due to ERM concentration with rate $1/\sqrt{n}$ going for $1/\sqrt{n}$ error may be "good enough".

Remark 2:

The above bounds are all in expectation
Could we improve them?

Simple idea: Use Markov's Ineq.

$$\Pr(|X| > a) \leq \frac{E[X]}{a}$$

## Remark 3:

GD is trivially parallelizable
SGD is inherently serial!

How could we parallelize?

## Next week:

Tue Lecture : SVRG

Thu. Lecture: RCD + importance sampling.